DOCUMENT RESUME

ED 093 900 TM 003 715

AUTHOR Smith, Donald M.
TITLE Determining Learning Sequences from a Difficulty-Scaling of Test Items.
PUB DATE [Apr 74]
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS *Achievement Tests; Arithmetic; *Complexity Level; *Internal Scaling; *Item Analysis; Measurement Techniques; Statistical Analysis; *Test Construction; Test Reliability
IDENTIFIERS Instructional Sequencing; Monotonic Deterministic Test Model; *Scaled Achievement Tests

ABSTRACT

The concept of scaled achievement tests is discussed and a method of selecting those items of a test that form the most scalable (i.e., having the highest coefficient of reproducibility) subset is presented. Sometimes called a monotonic-deterministic model, this type of test assumes that the test items may be sequentially ordered. To determine the probability of obtaining a given coefficient of reproducibility, two statistics are required: the expected value of the coefficient; and the standard error of the expected value. The procedures used to eliminate the items that do not fit the model are described by example. Tests were constructed to measure two skills, elementary addition and elementary subtraction, and were administered to pupils in grades 2-6. The resulting data were statistically analyzed and a coefficient of reproducibility was calculated to determine how well the tests had been scaled. The tests were then administered to another sample. The results of the replication were much the same: internal consistency estimates of test reliability were almost identical for both tests. Most significantly, it was demonstrated that both of the hypothesized orderings of tasks were stable across samples. (Author/RC)

26-07

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

TM 003 715 — ED 093900

DETERMINING LEARNING SEQUENCES FROM A DIFFICULTY-SCALING OF TEST ITEMS

Donald M. Smith
Ball State University

A paper presented at the annual meeting of American Educational Research Association, held in Chicago, April 15-19, 1974.

The purpose of this article is to briefly discuss the concept of scaled achievement tests and to present a method of selecting those items of a test that form the most "scalable" sub-set. "Most scalable" being used in the sense of having the highest coefficient of reproducibility (R_p).

The underlying assumption of this type of test, sometimes called a monotonic-deterministic model, is that the test items may be sequentially ordered. Under this assumption a person who has a score of "m" on an "n" item test will have correctly answered the first "m" items and incorrectly answered all remaining items. It is thus possible to reconstruct a person's entire response vector given only his total test score. While there are many possible advantages to this approach to testing it would seem that one, the determination of a "most" logical order in which a skill may be taught, is directly relevant to the topic discussed by this symposium.

Almost all of the reported work that has been done using this model has involved the measurement of attitudes and there is very little in the literature concerning it's application to achievement tests. Cox and Gordon (1966) developed a simple arithmetic achievement test based on the model: and Kohen-Rox (1967), using a variation of the Guttman procedure (H technique), reported very good results in his attempts to scale developmental sequences in infants. The studies in the area typically suffer from two defects: they scale items within a single sample and rarely attempt to replicate their findings on another sample, and they fail to provide an estimate of the statistical significance of the obtained R_p's. And each of these is critical in studies of this nature. In the first case the obtained scaling may be a non-reproducible situational artifact and hence meaningless; while in the second case the obtained value of R_p, which is a representation of the degree to which the test items have been scaled, may not be statistically significant. Regarding this last

point most researchers are content to simply state that a value above .85 or .90 represents a significant R_p. Cox and Graham, for example, stated that a value of .90 was acceptable and concluded that their obtained value of .92 was significant.

The continued use of such unsupported figures is no longer defensible. Considerable work has been done in the area and numerous methods of estimating the significance of an obtained value of R_p have been proposed. The earliest of these, and the basis for all subsequent proposals, was presented by Goodman (1956). This procedure was later simplified and extended by Sagi (1959) for one method of computing R_p. Sagi's work was further simplified and generalized by both Castellan (1969) and Cotton (1969). It is now possible, based on the above research, to determine the significance of an obtained value of R_p that has been computed by the method proposed by Goodman.

Using the procedures developed by the cited researcher's it is now possible to determine the probability of obtaining a given R_p. Two statistics are required for this evaluation: the expected value of the coefficient (and it should be noted that the minimum value for a test consisting of dichotomously scored items is .75); and the standard error of the expected value.

Since the underlying rational for a scaled test is that successful performance on any given item is dependent upon successful performance of the preceding item, the first step in devising a method of scaling items is to obtain a statistic that will estimate the statistical significance of the differences between the difficulty indices (DI) of the items that make up the test. An appropriate statistic for this purpose is the standard error of the difference between correlated proportions.

The formula is used to construct a k x k matrix (Z) whose elements are standard deviates (z scores) and represent the "distance" between each pair

of test items. Each element in the matrix is computed by dividing the difference between the corresponding pair of difficulty indices by it's associated error term.

Table 1 contains the upper half of such a matrix formed from the ten items of an experimental test in division.

Having formed the matrix the next step is to determine which of the "k" test items fit the model. In simple terms this involves the elimination of those items whose vector contains a value that is less than the value associated with the chosen level of significance. The procedure that is used to eliminate the items that do not fit the model can best be described by example and for this purpose the data presented in Table 1 will be used.

1. Choose the level of random error that is permissible and obtain the z score that corresponds to the one-tailed probability. In this case alpha will be equal to 0.05 and the corresponding z score is 1.64.
2. Locate the item in the matrix that has the highest difficulty index (i.e., the easiest item). In the example this is item one.
3. Check each element in row one to see if one of the values are greater than one minimum value (1.64). They are, so a 1 is placed above column one. This is the base item.
4. Locate the lowest value in the base row (1). In the example this is 1.73 and is in column four. Check row four to see if all of it's elements are greater than the minimum value. They are so a 2 is placed above column four.
5. Locate the next lowest value in the base row: 3.83 in column five. Column five is checked and one entry, that in row two, is less than the minimum value. Column two is then checked to see if it contains

TABLE 1

Matrix of Z Scores Representing the Differences Between Ten Correlated Proportions

	1	2	3	4	5	6	7	8	9	10
1	.857	4.56	7.64	1.73	3.83	8.78	10.78	10.27	8.63	11.24
2	-	.746	5.35	3.00	0.15	6.64	9.55	8.69	6.74	10.06
3	-	-	.582	6.93	5.25	2.71	7.04	5.84	2.15	7.31
4	-	-	-	.820	3.05	8.56	10.45	9.83	8.39	10.91
5	-	-	-	-	.750	7.28	9.60	8.74	6.81	9.93
6	-	-	-	-	-	.504	4.99	3.71	0.13	5.42
7	-	-	-	-	-	-	.357	1.29	4.45	1.34
8	-	-	-	-	-	-	-	.393	3.61	2.83
9	-	-	-	-	-	-	--	-	.508	6.34
10	-	-	-	-	-	-	-	-	-	.316

NOTE: Item difficulties are given in the diagonals.

TABLE 2

Reduced Matrix Containing Those Items Having no Entry Less Than 1.64

	1(1)	2(4)	3(2)	4(3)	5(6)	6(8)	7(10)
1	.857	1.73	4.56	7.64	8.89	10.27	11.24
2	-	.810	3.00	6.93	8.56	9.83	10.91
3	-	-	.746	5.35	6.64	8.69	10.06
4	-	-	-	.582	2.71	5.84	7.31
5	-	-	-	-	.504	3.71	5.42
6	-	-	-	-	-	.393	2.83
7	-	-	-	-	-	-	.315

NOTE: Original items numbers are in parentheses.

more than one value that is less than the minimum value. It does not so the entries in each of the columns are summed. The sum of column two is 54.74 and the sum of column five is 54.64. Since column two has the greatest average z this item is retained and a 3 is placed above column two. A line is drawn through row and column five. Had the number of lesser entries contained in the two columns been unequal, the column containing the greater number of lesser entries would have been eliminated. This is the second of two methods that can be used to eliminate items that do not fit the model.

6. The next lowest value in row one is that contained in column three. Since all of the entries in column three are greater than 1.64 a 4 is placed above column three.
7. Column nine contains the next lowest value in row one. There is one entry in column nine that is less than 1.64 so the associated column, six, is checked and found to also contain one value that is less than 1.64. The sum of column six is equal to 40.94 and the sum of column nine is equal to 40.44: item nine is therefore eliminated and a line is drawn through row and column nine.
8. Column eight is next in line and one entry, that in row seven, is less than the minimum value. Checking column seven it is found to contain two entries that are less than 1.64. The item is deleted, a line is drawn through row and column seven, and a 6 is placed above column eight.
9. The last column to be checked is column ten, which contains no entry

that is less than 1.64. A 7 is placed above column ten and the elimination of the items that do not fit the model is completed.

It has been possible, in this example, to retain seven of the original items and use them to construct a test that should meet the assumptions of the monotonic-deterministic model to a high degree. The reduced matrix, with the original item numbers given in parenthesis, is presented in Table 2. The difficulty indices of the retained items are contained in the diagnosis of the matrix.

The seven items that were retained have an average difficulty index of .6025, a standard deviation of .249, and a pronounced positive skew. This last is of importance since one of the criticisms most often leveled at tests based on this model is that they can be constructed only if there are large differences in the difficulty indices of adjacent items (Nunnally, 1967, pp. 64-66). This example would seem to offer partial refutation to these criticisms, since five of the items have difficulty indices greater than .5 and there are differences in the difficulty indices of adjacent items as small as .0369 ($D_1 - D_2$) and the average difference between adjacent items is .0898.

The procedure described above has proved useful in several different situations. The author has used it to scale two achievement tests in arithmetic and the results, obtained in grades 2-6 in one school system, were replicated in another school system. The results of this study, as well as those from a later replication in kindergarten, are given in the second part of this paper. Although the procedure was designed primarily for the scaling of items within a test it may also be used to scale other types of stimuli. For instance,

King (1972) has used it to scale reading passages in the elementary grades and Huntley (1971) has used it to scale tests representing the levels of Gagne's hierarchy of learning.

Construction, Analysis and Cross-Validation of the Scales

The second part of the study involved three separate tasks: the construction of the scales; the administration of these to a standardization sample and the analysis of the resulting data; and the subsequent administration of the resultant scales to a second sample, in order to determine if the obtained results were stable. These three steps are described in the following sections.

Construction of the Scales

It was decided to construct tests to measure two very simple skills: elementary addition and elementary subtraction. Although any success that might be attained using such simple skills would not serve as proof that the method would be successful if applied to other, more complex skills, failure on these simple tasks would strongly indicate that the basic concept of scaled achievement tests is impractical.

A listing of the hypothesized tasks involved in addition and subtraction was compiled and arranged in the expected order of difficulty. There were 20 tasks in the addition list and 23 in the subtraction.

The listing was constructed by starting with the simplest possible task in the skill and defining the next task as that which required the smallest possible increase in knowledge. For example, in addition the easiest possible task is to add two 1 digit numbers together to obtain another 1 digit number. The smallest possible increment in required knowledge would be the addition of

a third digit while retaining the constraint that the sum be less than 10. The process was continued until no further meaningful increments could be made.

The basic concept of this approach is that it is possible to determine whether an individual can perform a stated task with only a specified amount of random error. This requires that we have several samples of a person's performance on the task rather than the typical single sample of performance. The number of samples required will be determined by the amount of random error that the experimenter is willing to accept. This was set at the 0.05 level for this study and, since it was desired to have four-alternate multiple choice items, it was decided that four items would be a sufficient sampling of the behavior. Under these conditions correct answers to three or four of the items would be accepted as an indication of "mastery" of the sampled skill ($p = .0547$).

Experimental forms of both the addition and subtraction test were constructed: the former containing 4 similar items from each of its hypothesized tasks, for a total of 80 items; and the later containing 4 similar items from each of its hypothesized tasks, for a total of 92 items.

Analysis of the Data from the Standardization Sample

The tests were administered to the pupils in grades 2 through 6 in two elementary schools in Wakulla County (Florida). Both are rural schools and, although both are integrated, one school (Shadeville) is primarily black while the other (Sopchoppy) is primarily white. The number of pupils who took each of the tests, broken down by grade and school, is contained in Table 3.

TABLE 3

Description of the Wakulla Sample by Grade, School and Test

	Addition			Subtraction		
Grade	Shade-ville	Sopchoppy	Both	Shade-ville	Sopchoppy	Both
2	22	19	41	23	24	47
3	27	26	53	25	28	53
4	29	24	53	28	28	56
5	22	20	42	21	21	42
6	30	26	56	30	19	49
Total	130	115	245	127	120	247

Each test served as its own answer sheet. This method was chosen over separate answer sheets, even though it required more time to score the tests, since research (Cashen & Ramseyer, 1969) indicates that there is a marked and significant lowering of scores when separate answer sheets are used earlier than the 4th grade.

Two matrices of scores were created: $\underset{\sim}{A}$, the matrix for the addition test, being n_a x 20 and $\underset{\sim}{S}$, the matrix for the subtraction test, being n_s x 23. Each element in a matrix represents the number of items that were correctly answered on a given task. For example:

$$a_{ij} = \sum X_{ijk} \quad \text{where } k = 1,2,3,4$$

represents the score of the ith person on the jth task of the addition test. The obtained matrices were then transformed by substituting a 1 for each element that was ≥ 3 and a 0 for those that were < 3.

These matrices were then analyzed, using the procedures previously described and 9 items from each of the two experimental forms were found to meet the requirement, a minimum z of 1.64, for inclusion in the final test. The resulting matrices are presented in Table 4. The tasks corresponding to the 9 items, for each of the tests, are given in Table 5.

It was necessary, in order to compute the coefficient of reproducibility, to determine a total score for each subject. This was defined as the number of the first item that preceded the second zero and which contains a 1. This was an arbitrary choice on the part of the experimenter and several other scoring methods (i.e., the sum of the 1's, the sum of the 1's that preceded the first 0, etc.) could have been chosen. To illustrate how the method works consider 3 subjects, A, B, & C, who have the following items scores:

	Items								
	1	2	3	4	5	6	7	8	9
Subject A	1	1	0	1	1	1	0	0	1
Subject B	1	1	0	0	1	0	0	0	0
Subject C	0	1	0	0	0	0	0	0	0

Subject A would receive a total score of 6: this being the number of the last item preceding the second zero that was scored as a 1. Subject B receives a total score of 2, as does Subject C.

Since there was only a single administration of the two tests the only reliability coefficients that could be computed were estimates based on the internal consistency of the tests. Table 6 gives a complete listing of the obtained estimates of test reliability.

The obtained coefficients were acceptable and, considering the number of items in the tests, the small sample sizes in the individual classroom estimates, and the restricted range within classrooms, most of them could perhaps be called large.

TABLE 4

Matrix of Z Values Representing the Differences Between the Difficulty Indices of the 9 Items in the Two Tests

	1	2	3	4	5	6	7	8	9
1	-	3.33	5.51	7.39	8.47	9.62	10.29	11.19	12.53
2	4.71	-	3.57	4.92	7.21	8.34	9.36	10.44	11.71
3	6.41	2.34	-	2.55	4.82	6.41	7.74	9.04	10.56
4	8.21	4.71	2.97	-	2.56	5.00	6.53	7.89	9.76
5	8.89	6.11	4.53	2.06	-	3.06	4.57	6.22	8.37
6	10.25	7.63	6.64	4.12	2.32	-	2.19	4.34	7.04
7	11.49	9.48	8.40	7.32	5.78	3.57	-	2.67	5.66
8	12.49	10.17	9.26	8.01	7.04	5.25	2.79	-	4.52
9	13.04	11.62	10.70	9.56	8.55	7.37	5.37	3.28	-

NOTE: Entries above the diagonal are for the additional test.
Entries below the diagonal are for the subtraction test.

	Item Difficulties									Mean
Add	.8694	.7878	.6980	.6245	.5469	.4612	.4000	.3347	.2122	.5493
Sub.	.8450	.7171	.6589	.5620	.5039	.4302	.3333	.2674	.1861	.5004

TABLE 5

Task Descriptions

Addition

Task 1 - add two 1 digit numbers to obtain a 1 digit number.

Task 2 - add three 1 digit numbers to obtain a 1 digit number.

Task 3 - add three 1 digit numbers to obtain a 2 digit number that is greater than or equal to 20. Carrying is required.

Task 4 - add two 2 digit numbers to obtain a 2 digit number. No carrying is required.

Task 5 - add two 2 digit numbers to obtain a 2 digit number. Carrying is required.

Task 6 - add three 2 digit numbers to obtain a 2 digit number. Carrying is required.

Task 7 - add three 3 digit numbers to obtain a 4 digit number. Carrying is required in all 3 columns.

Task 8 - add two 4 digit numbers to obtain a 5 digit number. Carrying is required in the 1st two columns.

Task 9 - add 4 numbers, one each of 2, 3, 4 and 5 digits, to obtain a 5 digit number. Carrying is required in the 1st 4 columns.

Subtraction

Task 1 - subtract a 1 digit number from another 1 digit number to obtain a 1 digit number.

Task 2 - subtract a 1 digit number from a 2 digit number to obtain a 2 digit number. No carrying is required.

Task 3 - subtract a 1 digit number from a 2 digit number to obtain a 1 digit number. Carrying is required.

Task 4 - subtract a 2 digit number from a 2 digit number to obtain a 2 digit number. No carrying is required.

Task 5 - subtract a 2 digit number from a 3 digit number to obtain a 3 digit number. No carrying is required.

Task 6 - subtract a 1 digit number from a 3 digit number to obtain a 3 digit number. Carrying is required.

Task 7 - subtract a 2 digit number from a 2 digit number to obtain a 1 digit number. Carrying is required.

Task 8 - subtract a 2 digit number from a 3 digit number to obtain a 3 digit number. Carrying is required in the 1st 2 columns.

Task 9 - subtract a 4 digit number from a 4 digit number to obtain a 3 digit number. Carrying is required in all columns.

The next step was to determine how well the tests had been scaled. The usual procedure is to calculate a Coefficient of Reproducibility (R_p) and use this figure as an index of the ability to reproduce a subject's item responses given only his total score. The computation of the coefficient is a simple matter, requiring only 3 operations per subject (Goodman, 1956).

The main problem with Coefficients of Reproducibility has been, until recently, determining the significance of an obtained value. Guilford (1954, p. 461) has suggested that the obtained value of R_p should always be greater than .85, and preferably greater than .90, to be considered significant. However, recent research in the area by Castellan (1969), Cotton (1969), Goodman (1956) and Sagi (1959) has produced methods that make it possible to determine the exact probability of obtaining a given value. Two steps are required: (a) computing the lower bound of the coefficient for the sample R_p; and (b) computing the standard deviation of the sample value (S_{R_p}). The formula for the lower bound of the coefficient is:

$$(1) \qquad R_p' = 1 - \sum_{i=1}^{k-1} \frac{p_i q_{(i+1)}}{k}$$

where: p_i = probability of a correct response on the ith item

q_i = probability of an incorrect response on the jth item

k = number of items in the test

As can be seen, the value in the numerator will be maximized when all $p_i = .5$ at which time:

$$(2) \qquad R_p' = 1 - \frac{.25\,(k-1)}{k}$$

which will give values of R_p' that range from a upper limit of .875 for a 2 item test to a lower limit of .75 for a test having an infinite number of items. For a 9 item test R_p is .7778 when all p_q's are .5. However, when both these tests have p_q that range from .2 to .8 we can expect their R_p' to be considerably higher. Goodman (1956) has shown that the statistic

(3) $$\frac{R_p - R_p'}{S_{R_p}}$$

has a distribution that approximates normality with (0,1). The expression in the denominator is the standard deviation of R_p and is equal to

(4) $$S_{R_p} = \sqrt{\frac{\sum_{i-1}^{k-1} S_i^2 \left(S^2_{(i+1)}\right)}{nk^2}}$$

where: S_i^2 = variance of the ith item and all other terms are as defined previously.

Using these two statistics it is possible to estimate the significance of any obtained value of R_p; although, as Castellon (1969) points out, a difference as small as .025 between the R_p and the R_p' will be significant at the .05 level whenever the difference is obtained from a test having 10 or more items that was administered to 30 or more subjects.

Coefficients of Reproducibility were computed for schools, classrooms and classrooms within schools and the obtained values are given in Table 7. All coefficients were significant at or beyond the .05 level.

Analysis of covariance was performed, using the absolute deviation of each subject's total test score from 4.50 as the covariate, on the obtained

values of R_p to see if there were any differences between the values obtained by the different schools, the different classrooms of the different classrooms within schools. The covariate was employed in order to partial out the portion of the obtained coefficient that could be attributed to a person's total score; for, as discussed earlier, R_p is a direct function of the average difficulty level of the items within a test, which in turn determine the average total score obtained by the subjects. In other words, we may expect, by chance alone, to obtain higher values of R_p on those tests that have either an extremely high or an extremely low average difficulty index. It would have been possible to have used the R_p or the $\bar{D}$ of a group as the covariate - however, since these would have been the same for all members of the group, the absolute deviate was deemed the better choice.

The results of these analyses are given in the two source tables contained in Table 8. There were no significant differences attributable to either schools, grades or schools by grades; and, oddly enough, only in the subtraction test was the covariate significant.

The Cross-Validation (Moore) Sample

This portion of the study was carried out in order to determine if the obtained ordering of tasks was stable across settings. The school chosen was Moore Elementary, a school within the Leon County (Tallahassee) Florida School District. Moore is a new school and has been integrated since it was first opened in 1966.

Since it was desired to obtain information concerning the test-retest and alternate forms reliabilities, an alternate form of each of the two tests was constructed. This was accomplished by writing 4 new items for each of

TABLE 6

Internal Consistency Estimates (KR-20)
of Test Reliability

	Addition			Subtraction		
Grade	Shade-ville	Sopchoppy	Both	Shade-ville	Sopchoppy	Both
2	.7737	.7888	.7901	.7771	.6928	.7618
3	.7647	.7435	.7662	.8340	.7789	.8062
4	.9363	.9007	.9375	.9352	.8910	.9244
5	.9087	.8147	.8891	.8999	.8976	.9036
6	.8720	.9371	.9162	.9034	.9002	.9059
All	.9062	.9997	.9073	.9304	.8950	.9177

TABLE 7
Obtained Coefficients of Reproducibility and
Associated Values of Z

	Addition			Subtraction		
Grade	Shade-ville	Sopchoppy	Both	Shade-ville	Sopchoppy	Both
2	.9516	.9415	.9496	.9227	.8982	.9102
	17.30	14.51	15.91	9.00	14.59	11.44
3	.9603	.9573	.9589	.9487	.9008	.9239
	25.00	23.44	24.23	10.09	12.22	14.31
4	.9333	.9213	.9267	.9226	.9524	.9352
	10.75	10.00	10.38	11.97	14.39	13.17
5	.9444	.9557	.9497	.9630	.9153	.9392
	10.05	11.46	10.65	12.34	8.09	10.31
6	.9606	.9402	.9513	.9148	.9474	.9347
	14.07	11.51	12.52	6.74	10.42	8.85
All	.9514	.9430	.9487	.9346	.9216	.9323
	30.20	27.86	28.97	25.75	25.70	25.72

TABLE 8

Source Table for the Analysis of Covariance on the
Coefficients of Reproducibility Obtained

		Addition Test		
Source	df	SS	MSS	F
Mean	1	188.99963		
Schools	1	.00003	.00003	.005
Grades	4	.04188	.01047	2.292
S x G	4	.01065	.00266	.582
Covariate	1	.00243	.00243	.532
Error	234	1.06107	.00453	
Total	245	190.00000		
		Subtraction Test		
Mean	1	216.67032		
Schools	1	.00075	.00075	.079
Grades	4	.02458	.00615	.653
S x G	4	.08800	.02200	2.332
Covariate	1	.31263	.31263	33.327
Error	236	2.20000	.00939	
Total	247	219.00000		

the 9 tasks that are sampled by each of the tests (a total of 72 new items) and placing these, along with the original 72 items, in an item pool. Four items from each task were drawn at random from the pool and placed in Form A, and the remainder in Form B, of the appropriate test. Although each of the new forms contains items taken from the original test they are both different from the form that was administered to the Wakulla sample.

A counter-balanced design was used in order to determine what proportion of the test variance could be attributed to the various treatment effects: sequence, test-retest vs alternate forms, grades and the interactions. This procedure required that each subject be administered the test twice. One-half of the subjects in a classroom took Form A on the first administration and the other half taking Form B. On the second administration one-half of the subjects received the same form that they had taken on the first administration and the other half received the alternate form. This meant that a quarter of each classroom, or as close to that proportion as classroom size permitted, took the tests in the following sequence: Form A - Form A; Form A - Form B; Form B - Form A; and Form B - Form B. The tests were given by the experimenter and another graduate student. The first administration was on a Tuesday morning and the second was on the following Thursday morning. The pupils marked their answers directly in the test booklets and these were later handscored by the writer. A breakdown of the number of different pupils in each classroom that received each of the four testing sequences is given in Table 9.

TABLE 9

Number of Subjects That Received Each Test Sequence

Grade	AA	AB	BA	BB	TOTAL
2	6	5	8	6	25
3	6	7	6	7	26
4	6	7	7	5	25
5	5	6	7	5	23
6	6	5	3	5	23
Total	29	29	30	34	122

The results of two analyses's of covariance, using the score obtained on the first test administration as the covariate, indicate that there are no significant differences in the adjusted mean scores obtained by the various treatment groups on the second administration. These results are given in Table 10.

Since there were no significant differences it is permissible to compute pooled estimates of the reliability of the two tests: these were .8533 for the addition test and .8678 for the subtraction test. These values are slightly lower than the internal consistency estimates of test reliability which are given in Table 11.

The obtained values are in very close agreement--the largest difference being - .0082 between the coefficient obtained for the two administrations of Form A of the addition test. There is also very close agreement between these values and those computed (see Table 6) for the Wakulla sample.

The next, and most important, step in the analysis was to determine if the ordering of tasks hypothesized from the Wakulla sample had been obtained. Since there had been no significant effects introduced by the various treatment groups the matrix of Z scores was computed using the pooled data from the two administrations. This gives an effective sample size of 244 for each test. The values obtained and the difficulty indices of the items are given in Table 12. The entries above the diagonal are for the addition, and those below the diagonal are for the subtraction test.

The mean value of the addition test is some 25% higher than that obtained for the Wakulla sample while the mean value of the subtraction test differs from that obtained by the Wakulla sample by less than 6%. The hypothesized

TABLE 10

Source Tables for the Analysis of Covariance Performed on the Test Scores Obtained on the Two Administrations

Source	df	SS	MSS	F
		Addition		
Mean	1	4679.123		
Treatments	19	58.913	3.101	1.457
Grades	4	16.642	4.161	1.938
TRT-Alt	1	7.077	7.077	3.324
Form A First	1	.027	.027	.013
Form B First	1	.266	.266	.125
TRT-A x Grades	4	8.649	2.162	1.008
Form A First x Grades	4	10.225	2.556	1.189
Form B First x Grades	4	7.833	1.948	.911
Covariate	1	731.923	731.923	383.696
Error	101	215.041	2.129	
Total	122	5685.000		
		Subtraction		
Mean	1	3020.074		
Treatments	19	46.834	2.465	.809
Grades	4	11.126	2.782	.913
TRT-Alt	1	.053	.053	.018
Form A First	1	1.287	1.287	.422
Form B First	1	4.943	4.943	1.622
TRT-A x Grades	4	9.654	2.411	.798
Form A First x Grades	4	4.118	1.030	.338
Form B First x Grades	4	3.453	.863	.283
Covariate	11	1080.268	1080.268	307.699
Error	101	307.824	3.048	
Total	122	4455.000		

TABLE 11

KR-20 Internal Consistency Estimates of Test Reliability

	Addition Test		Subtraction Test	
Administration	Form A	Form B	Form A	Form B
First	.8924 (59)	.8975 (63)	.9275 (59)	.9276 (63)
Second	.9002 (63)	.8929 (59)	.9340 (63)	.9303 (59)

ordering of tasks for the addition test was reproduced perfectly and there was only one minor reversed, between items 3 and 4 in relation to item 7 on the subtraction test. This is a very minor difference and is not deemed to be a serious defect in the replication.

The values of R_p that were obtained for the various grades, administrations and grades by administrations, are given in Table 13. These were higher than those obtained by the Wakulla sample.

The results obtained from this replication of the experiment were much the same as those that had been obtained in the original study. The KR-20 internal consistency estimates of test reliability were almost identical for both of the tests; and, although the values of R_p obtained in the replication were higher than those in the original study, this cannot be considered a defect. The most significant result of the replication was the demonstration that both of the hypothesized orderings of tasks was stable across samples, the importance of which has already been covered.

General Discussion

The primary goal of the study, to construct two simple sealed achievement tests in arithmetic, was attained and the obtained scales proved to be stable over groups. Although these results cannot serve as proof that it will be possible to construct scalable orderings of the tasks involved in higher order skills it does indicate that it may be possible to do so for those skills whose component tasks can be precisely defined.

The benefits of such tests to education could be quite large. At the individual level it would be possible to use the test as a diagnostic instrument since each task requires the knowledge of one operation not needed to perform

TABLE 12

Matrix of Z Scores Representing the Significance of the Difference Between the Difficulty Levels of the 9 Items in the Two Groups

	1	2	3	4	5	6	7	8	9
1	-	1.00	4.95	6.14	8.63	9.43	10.54	10.95	11.62
2	6.01	-	4.49	5.77	8.56	9.27	10.39	10.82	11.49
3	6.79	1.67	-	2.14	6.31	7.14	8.80	9.29	10.16
4	6.85	2.04	.16	-	5.18	6.51	7.90	8.56	9.39
5	7.39	3.32	1.77	2.29	-	1.64	4.43	5.35	6.62
6	8.78	5.53	4.26	4.87	2.97	-	3.67	4.52	5.84
7	9.91	7.76	6.65	7.02	5.94	4.00	-	1.66	3.70
8	10.63	8.55	7.41	7.88	7.15	5.25	2.12	-	2.24
9	11.15	10.12	9.99	9.59	8.77	7.09	5.13	2.61	-

NOTE: Values for the addition test are above the diagonal
Values for the subtraction test are below the diagonal

	Difficulty Indices of the 9 Items in the Two tests									Mean
D_a	.9528	.9329	.8186	.7788	.6199	.5779	.5098	.4563	.4045	.671
D_s	.8170	.6632	.6201	.6176	.5711	.4908	.4028	.3548	.2427	.5311

TABLE 13

Values of R_p and the Associated Z Scores Obtained for the Moore Sample

	Addition			Subtraction		
Grade	1st Admin.	2nd Admin.	Both	1st Admin.	2nd Admin.	Both
2	.9861 24.12	.9722 17.37	.9796 20.03	.9954 17.03	.9815 14.55	.9887 15.29
3	.9671 19.39	.9506 16.39	.9589 17.81	.9630 15.34	.9712 15.86	.9671 15.61
4	.9511 15.03	.9378 13.13	.9444 14.12	.9556 16.47	.9467 13.50	.9511 14.87
5	.9420 8.03	.9710 11.08	.9565 9.37	.9565 12.22	.9710 13.80	.9638 12.90
6	.9710 6.09	.9807 9.43	.9759 7.84	.9565 9.09	.9710 12.25	.9638 10.67
To[illegible]	.9636 31.04	.9622 30.48	.9628 30.76	.9654 29.71	.9681 29.92	.9668 29.84

the preceding task. It would even be possible, by examining the responses made to those items that were failed, to determine the nature of the errors that were being made. In addition, by the use of the proper statistical techniques, it might be possible to ascertain what minimum level in one skill is required to be able to successfully acquire another allied, but more complex, skill.

Since the difference between the entry score and the exit score is an exact statement of the tasks that have been learned during a given period of time it would be possible to make more accurate evaluations of the relative effectiveness of schools, programs, methods of instruction and teachers.

Another advantage of this approach is the reduction in the overall amount of time that would be required for testing; since, having once determined the highest task that a child can successfully perform, we can begin all future test administrations with the last task that was successfully passed in the previous administration. A final advantage of the method is the case with which parallel forms can be constructed. Given the descriptions of the tasks that are to be measured and the nature of the incorrect alternate responses it should be possible for any classroom teacher to produce an infinite number of different, but equivalent, forms. This would also serve as a check on how well the tasks had been described and could easily be checked by having two teachers construct forms of the test and then administer both tests to the same sample. The correlation between the two test scores, corrected for attentuation, would be an indication of the degree to which the tasks had been defined (Cronbach, 1969, p.44).

The proposed methodology is very flexible and is capable of constructing

stable orderings of items that have difficulty levels much closer than was the case in this study. The determining factor is the amount of Type II error that the experimenter is willing to accept.

Additional research in the area, designed so as to determine the applicability of the procedure to the scaling of higher order skills, appears to be warranted.

REFERENCES

Bryden, M. P. "A Non-Parametric method of item and test scaling." Educational & Psychological Measurement, 1960, 20. 311-315.

Cashen, Valjean M. & Gary C. Ramseyer. "The use of separate answer sheets by primary age children." Journal of Educational Measurement, 1969, 6, 155-158.

Castellan, N. John, Jr. "A note on the expected value of a coefficient of reproducibility." Multivariate Behavioral Research, 1969, 4, 363-369.

Cotton, John W. "The sensitivity of coefficients of reproducibility to intercorrelations among items, dispersion of pass-fail proportions, and multiple factor structure." Multivariate Behavioral Research. 1969, 4, 346-361.

Cox, Richard C. & Glenn T. Graham. "The development of a sequentially scaled achievement test." Journal of Educational Measurement, 1966, 3, 147-150.

Goodman, L. A. "Simple statistical methods for scalogram analysis." Psychometrika, 1956, 21, 179-188.

Green, Bert F. "A method of scalogram analysis using summary statistics." Psychometrika, 1956, 21, 79-88.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Guttman, Louis. "A basis for scaling qualitative data." American Sociological Review, 1944, 9, 139-150.

Guttman, Louis. "The basis for scalogram analysis," in S. A. Stouffer (ed.), Measurement and Prediction. Princeton, N.Y.: Princeton University Press, 1950.

King, F. J., Jr., Department of Educational Research, Florida State University, Tallahassee, Florida. Personal communication, 1972.

Kohen-Raz, R. "Scalogram analysis of some developmental sequences of infant behaviors as measured by the bayley infant scale of mental development." Genetic Psychological Monographs, 1967, 76, 3-21.

Menzel, Herbert. "A new coefficient for scalogram analysis." Public Opinion Quarterly, 1953, 17, 268-280.

Nunnally, Jum C. Psychometric theory. New York: McGraw-Hill, 1967.

Sagi, P. C. "A statistical test for the coefficient of reproducibility." Psychometrika, 1959, 24, 19-27.

Smith, Donald M., Construction and validation of two scaled achievement tests in arithmetic. Mimeo, Department of Educational Research, Florida State University, Tallahassee, Florida, 1971.

Stouffer, Samuel A., et al. "A technique for improving comulative scales." Public Opinion Quarterly, 1952, 16, 273-291.

White. Benjamin W. "Measurement of reproducibility." Psychological Bulletin, 1957, 54, 81-99.